Introducing a Romanian Frequency List and the Romanian Vocabulary Levels Test

Csaba Z Szabo
Faculty of Languages and Applied Linguistics
The Open University
UK
csaba.szabo@open.ac.uk

INTRODUCING A ROMANIAN FREQUENCY LIST AND THE ROMANIAN VOCABULARY LEVELS TEST

(Abstract)

Vocabulary is considered essential to language learning, thus English word lists and tests based on frequency information have become the centre of attention for researchers, teachers and learners alike. As a result, it is argued hereby that frequency based word lists and tests should be adapted and regarded as key elements for teaching and learning Romanian as an additional language as well.

Since there are currently no reliable frequency lists and lexical tests in Romanian, this paper aims to bridge this gap by introducing the first Romanian Word List and the Romanian Vocabulary Levels Test. The list contains the 10,000 most frequent Romanian words and is based on the Romanian Balanced Annotated Corpus (ROMBAC, Ion, Irimia, Ștefănescu, Tufiș 2012).

The primary objective of the paper is to elaborate on the compilation criteria, the challenges involved and the benefits of such a list in the case of teaching, learning and curriculum design for Romanian as an additional language. The secondary objective is to present a practical application of the word list by introducing an exemplary Romanian lexical test, the Romanian Vocabulary Levels Test and examine its reliability and validity.

editura universității din bucurești®, 2015

Please cite as:

Szabo, Cz. (2015). Introducing a Romanian Frequency List and The Romanian Vocabulary Levels Test. *Current Issues in Linguistic Variation: The 14th international conference of the Department of Linguistics*, Vol. 2, University of Bucharest: Bucharest University Press, 304 – 316.

INTRODUCING A ROMANIAN FREQUENCY LIST AND THE ROMANIAN VOCABULARY LEVELS TEST

CSABA ZOLTAN SZABO
The Open University, UK

1. Introduction

This paper addresses the importance of frequency information and lexical tests in the case of teaching, testing and learning Romanian as an additional language. Since vocabulary is considered essential to language learning, word lists and tests based on frequency information have been widely used in English second language research and teaching (see Brezina and Gablasova 2015; Nation 2006; Treffers-Daller and Milton 2013). However, it is argued that despite the fact that vocabulary acquisition occupies a focal position in the Romanian language teaching context, a reliable Romanian frequency list and lexical test still do not exist.

The need for lexical development in and outside the classroom in the case of additional language learners has been recognised by teachers for decades and the foundation for this probably lies in the often repeated argument by Wilkins: "Without grammar very little can be conveyed, without vocabulary nothing can be conveyed" (Wilkins 1972: 111). Researchers and language teachers also acknowledge the fact that learners instinctively recognise the importance of lexical knowledge. For instance, Meara (1980: 221) claims that "most learners identify the acquisition of vocabulary as their greatest single source of problems". Furthermore, Schmitt (2010) argues that students when abroad refer to dictionaries rather than grammars. Therefore, finding the best methods to ensure learners make the most of their time and effort to increase lexical knowledge is a challenging task.

Vocabulary can be selected for classroom instruction in a number of ways and some of these are purportedly more successful than others. O'Laughlin (2012) and Schmitt and Schmitt (2014) found that many popular English coursebooks only cover a limited number of useful vocabulary items and do not provide enough repetition of lexical items.

Although dictionaries are essential tools for language learning and teaching, and their importance cannot be neglected, their success is dependent on the type of dictionary and the strategies learners employ. Using dictionaries to test vocabulary size or select targeted lexical items for instruction has not been entirely successful however. Nation and Coxhead assert that due to the dictionary spaced sampling method and the lack of a clear definition for the words used as standard counting units of measurement "almost all research on vocabulary size carried out during the twentieth century is virtually useless and at worst grossly misleading" (2014: 338).

Since the appearance of various digital corpora, lexical frequency information has been found to provide useful resources for teaching and testing languages. Based on reliable quantitative methods drawn on corpus linguistics, most researchers in the field agree on the fact that vocabulary can be divided into: *high frequency words* (i.e. the most frequent 2000

words), *general academic vocabulary* needed for the comprehension of academic texts and contexts, *technical/specialised vocabulary* (words frequently used in e.g. aviation or gastronomy) and *low frequency vocabulary* (i.e. the 'rest') (see e.g. Lessard-Clouston 2012 or Nation and Webb 2011). Among the reasons for this lies the fact that in any language a large amount of words is necessary for language use and frequency information can provide a realistic picture of frequent words that are used in general language or specific contexts. Moreover, as frequency lists and vocabulary tests are based on textual corpora, they represent authentic written and aural texts, making them even more attractive for both learners and teachers.

This study sets out to employ previously used methodologies in English to develop a frequency-based Romanian Word List and present a practical application of this word list in form of the Romanian Vocabulary Levels Test. This is justified by the fact that frequency lists for teaching Romanian as Second or Foreign language are non-existent despite their potential in a wide range of applications.

2. Teaching and learning vocabulary in the Romanian context

Over the past two decades or so a major shift in paradigm started to emerge in teaching Romanian to Hungarians and other nationalities whose first language is not Romanian. This paradigm shift is mainly characterised by the conceptualisation of language learning needs and the realisation that non-native speakers should be taught differently to native speakers of Romanian, by contrast to the traditional ways. This is also proved by the fact that teaching Romanian as an additional language gained gradually more attention as shown by a number of publications. Platon, Burlacu and Sonea (2011) in the *Procesul de predare-învățare a limbii române ca limbă nematernă (RLNM) la ciclul primar - The Process of Teaching-Learning Romanian as a Non-native Language in the Primary Cycle* (author's translation throughout) recognise this paradigm shift and highlight that a rich vocabulary is an essential parameter for decoding messages and can also increase the difficulty of a text. In conjunction with this, considering the Romanian teaching methods for monolinguals in the Republic of Moldova, Axenti and Verșina acknowledge the fact that teaching vocabulary in a systematic way allows learners to master the basic characteristics of word knowledge and "derivation with suffixes and prefixes has to be a means for enriching vocabulary" (Axenti and Verșina 2009: 89).

In comparison, Sîrghie (2009) also emphasizes that high-quality written and oral communication is characterised by mastering orthographic, orthoepic and punctuation rules and through continuous vocabulary development. Dina (2013: 1034) recognises vocabulary development as a key factor that is essential to the progression between the key stages of language learning and reports that once "the essential vocabulary" is mastered by learners, communication exercises are employed in order to further improve their lexical knowledge. The progress of newly learnt words from receptive to productive vocabulary is seen as an essential step in a number of publications (e.g. Axenti and Verșina 2009, Sîrghie 2009) however the means for selecting vocabulary for teaching, how 'essential' vocabulary is defined or the ways to evaluate students' progress remain unclear.

With regard to the first two issues (word selection and 'essential vocabulary') it has to be noted that Bârlea and Cerkez (2005), among many, make a distinction between *fundamental or basic vocabulary* ('*vocabular fundamental*', '*fond lexical principal*') and the 'rest' of vocabulary ('*masa vocabularului*'). The essential or basic Romanian vocabulary consists of

approximately 1500 words that are frequently used by "all language users" (p. 54), including body parts, colours, basic human actions, domestic and wild animals etc. By and large, it is questionable however, whether, for example, for a foreign speaker of Romanian on the verge of commencing his/her academic studies in Romania, words such as 'wolf', 'cherry' or 'ring' can be considered the sine qua non of language learning.

Using a more systematic approach, Biriș, Burlacu and Șoșa (2011) compiled a learner dictionary. Considered as the 'minimum vocabulary for Romanian', it "comprises 671 entries, 1410 pairs of antonyms and analogies, as well as over 2500 synonyms". They claim that "it represents an efficient means of lexical acquisition, since the antonyms are approached in a direct relation with synonyms and polysemous words" (cover page). In other words, the selection criteria for items included in this list are based on semantic characteristics (items must have either synonyms or antonyms), thus ignoring simple or abstract words that might not fit these categories, but could still be considered essential. Furthermore, there is a considerable amount of evidence to show that compared to thematic clustering of words, semantic clustering does not actually facilitate vocabulary acquisition due to the fact that the more distinct the words are (*eat and chocolate* vs. *blueberry and strawberry*) the easier is to learn them (see e.g. Erten and Tekin 2008; Tinkham 1993).

Regarding the third issue (evaluation of lexical development), Norel and Pop (2005) state that children's vocabulary develops considerably by the end of preschool. This idea springs from Golu, Zlate and Verza's (1992) Psychology manual for the end of high-school (year 11), which asserts that at around the age of 10-11, monolinguals know approximately 5000 words, most of which is part of their active vocabulary. It is hard to tell, however, how they define what a word is and how this has been measured.

It becomes obvious from the aforementioned that vocabulary and lexical development is recognised as a key concept in learning and teaching Romanian as an additional language in a wide range of documents. Although it is beyond the scope of the current study to verify or judge the validity or efficiency of these claims, they certainly raise a number of questions: What is it meant by 'essential' Romanian vocabulary and how can we define useful vocabulary based on contemporary tools and knowledge? What is the target vocabulary for Romanian language learners? How can vocabulary knowledge (a 'rich vocabulary') be measured at different stages?

3. English frequency lists and measuring vocabulary size

Learners engage with the target language in some way or from and in this process, they often meet highly frequent and some infrequent words. Many argue that high frequency words are encountered more often, thus the likelihood of mastering these words is greater as well. Consequently, learners' vocabulary knowledge will preponderate in the high frequency ranges, unless the target language or the input is highly specialised (e.g. aviation) (see e.g. Milton 2009; Nation 1983 or Schmitt 2010).

Since word frequency is quantifiable, there have been various attempts to estimate how many words are actually needed or known by users. Nation and Meara (2010) came to the conclusion that around 4-5000 word families (base word plus its inflections and derivatives) are required for an intermediate English proficiency and anything up to 9000 for advanced. Attempts have been made to establish the vocabulary size of native English speakers as well. Treffers-Daller and Milton (2013) review a number of studies in which educated native

speakers of English were reported to have vocabulary size estimations varying from 200,000 to as modest as 10,000 words. Their own investigation point to the conclusion that students' vocabulary size may well be at the more modest end of the spectrum and students' knowledge of approximately 10,000 words (entry level; 11,000 final year) shows a consistent variance around this figure (+/- 2,000). They also point out that despite students' limited range of vocabulary, the lexical scores obtained can be used to explain the variation in their academic performance: "students with larger vocabularies tend to score higher in their assignments and exams and to obtain higher degree classifications than those with smaller vocabularies" (Treffers-Daller and Milton 2013: 166).

In addition to vocabulary size, frequency data can be used to investigate the relationship between lexical knowledge and comprehension. The first 2,000 highly frequent English words, the General Service List (GSL) (West 1953), have been shown to provide 75-80% coverage of most texts. To put this differently, learners, who attain the most frequent 2,000 words of English, will encounter around 20 unknown words in 100 in a general English text and will comprehend nearly 95% of spoken English (Adolphs and Schmitt 2003). Nation (2006) investigated the requirements for the comprehension of English novels and newspapers, and found that learners with a lexicon covering the 8-9,000 most frequent words in the British National Corpus will have 98% lexical comprehension. Based on this and a large amount of empirical evidence, Schmitt and Schmitt (2014) recommend that *high frequency vocabulary* should be extended to the most frequent 3,000 words and below the 9,000 threshold the words should be categorised into a *mid-frequency vocabulary*. It is to be noted that these numbers mean word families and if these figures are translated into individual words, 8,000 families actually consist of over 34,500 individual words (Nation 2006).

It has been suggested that frequency information can be used to (1) set targets for students to acquire the essential coverage for understanding a wide variety of texts and (2) quantify their actual vocabulary size. Frequency lists can also provide an essential resource for achieving the necessary vocabulary in form of graded readers or mastering academic vocabulary in English for Academic Purposes contexts. Word lists such as the GSL and Coxhead's Academic Word List (AWL) (Coxhead 2000) have proved to be invaluable. The widespread use of these lists is also supported by the fact that Gardener and Davies (2013) recompiled the AWL (New Academic Vocabulary List) based on the COCA (Corpus of Contemporary American English; Davies 2008). Moreover, Brezina and Gablasova (2015), used four different language corpora to create a new GSL which features 2,122 core vocabulary items.

Another notable example is the JACET 8000 list (Aizawa 2006). This, slightly larger list is compiled from a large learner-oriented material, incorporates the majority of lexical items in other lists, more or less realistically distributes cognates and structure words and it is considered more suitable for speakers of Latin-based languages (Miralpeix 2008). Nation's BNC frequency list (2006) has been used for creating several vocabulary size tests. Schmitt and Schmitt (2014) compared it to the COCA and found that the first 9000 words provide coverage of just over 95% of this massive and diverse amount of data, which reiterates the importance of teaching high and mid-frequency vocabulary.

These realisations show that frequency lists and tests can provide a vast amount of information about the structure of a language, the targets for language learners, tools to reach these aims and quantitative and standardised tests for teachers and researchers to evaluate language learners' progress.

However, as Macoveiciuc and Kilgarriff (2010) pointed out Romanian is lacking a publicly available, large balanced corpus that would enable teachers and other stakeholders to improve teaching Romanian as an additional language. This is probably also the reason for the virtually non-existent Romanian graded readers (one example is the *First Romanian Reader for beginners: bilingual for speakers of English: 1*; Arefu 2014).

4. The Romanian Word List (RWL)

According to Nation and Coxhead (2014) to eliminate the difficulties represented by the dictionary sampling method and develop a vocabulary test, a suitable frequency list is essential. This is ideally derived from a contemporary textual balanced corpus that represents real language from a wide range of subjects distributed proportionally, and from authentic written and oral sources. Besides English (BNC, COCA etc.) and a handful of other languages (e.g. French), such corpus is hard to come by, especially in Romanian.

One notable example is the 50 million word RoWAC compiled by Macoveiciuc and Kilgarriff (2010), using the Web-as-Corpus method that can be accessed through the Sketch Engine (Kilgarriff et al. 2004). This was considered unsuitable for the purposes of this study as the sources mainly represent journalism the corpus cannot be regarded as a balanced corpus, and when it comes to web-based resources the texts' authenticity can also be questioned. Other corpora of Romanian are either restricted for the public or to journalism, or represent lexicographic corpora that cannot be used for gathering and establishing frequency information. This study uses the Romanian Balanced Annotated Corpus (ROMBAC; Ion, Irimia, Ștefănescu and Tufiș 2012). Contrary to what the name suggests, however, this corpus does not include oral texts, and the written texts are drawn from largely formal scholarly areas. Nevertheless, it is still the largest Romanian corpus (not web-resource based) available to date. According to the authors, discounting punctuation marks, the ROMBAC contains about 36,000,000 words evenly distributed into five genres: journalistic (news and editorials), pharmaceutical and medical short texts, legalese, biographies of the major Romanian writers and critical reviews of their works, and fiction (both original and translated novels and poetry). (Ion, Irimia, Ștefănescu and Tufiș 2012: 339).

The Romanian Word List has been developed from this corpus in two stages. At both stages the following rigorous adjustment criteria have been followed: punctuations, foreign words, numbers (including dates), proper nouns, abbreviations, duplications, and special and erroneous characters (e.g. %, ^, *) have been removed. Following this, the items on the list have been checked idividually to ensure that there were no oversights.

Based on Milton (2009) and Brezina and Gablasova (2015) it has been decided that the unit of measurement in the list will be lemmas instead of word families. Lemma or lemmas (lemmata) mean the headword (or stem) and the most common inflections without changing the part of speech. Thus *cook, cooks, cooking, cooked* belong to one lemma and *cooker and cookers* to a different one. According to Milton (2009), it is commonly accepted that employing lemmas as the basic unit of word counts is most practical and reliable as it draws on the fact that learners will master frequent derivations and inflections over the irregular or infrequent forms.

The first stage of developing the list involved using the lemmatized list that is provided with the corpus. However, once the raw list of lemmas which contained almost 500k items was adjusted using the above criteria, even amongst the most frequent lemmas it was possible to

find words from biology or medicine, such as *glicemie or ribavirină*. The word *pacient-patient* for example occurred over 36,000 times (frequency index) and thus it was the 27th most frequent word (rank) of Romanian. Due to the number of these words and the lack of possibility to filter them, these were left in the final list at stage one. However, these have been omitted during the word selection procedure for the RomVLT.

Stage two of the process was concerned with developing a more refined frequency list for Romanian. This required the five different sections of the ROMBAC to be revisited separately.

Table 1: The distribution of tokens and types in sections

Section	Tokens	Types	Percentage in ROMBAC
1 Journalism	1,922,109	50,945	7%
2 Literature	6,950,371	105,346	27%
3 Medical	6,783,005	362,782	26%
4 Legalese	6,269,543	248,354	24%
5 Biographies	3,716,031	223,592	14%
Total	**25,641,059**	**991,019**	**100%**

Table 1 shows the distribution of tokens (running words), in the present case lemmas, and types over the five different sections. The total number of lemmas extracted from the ROMBAC is over 25 million words (tokens) which amounts to just under 1 million different words (types). It should be noted that the medical section is at least 26% of the corpus and this indicates the amount of high frequency medical terms in the initial list. As this section (3) is considered highly specialized and it is unlikely that a general foreign language learner will account medical terms to this proportion, a decision was made to exclude this section to allow for more useful and general words among the highly frequent Romanian words. It was also decided to eliminate compound words (*floarea-soarelui – sunflower*) and frequent collocates (*dinainte de – before something; conform cu – according to*) from the frequency list. The reason for this was that frequent collocates are mainly built up by two highly frequent words or at least one (e.g. *de, pe, la, cu, din*), thus their presence would be duplicated in the list. Furthermore, their meaning can often be deduced from the individual words that make up the collocate. Additionally, due to the number of variations, regularities and frequency it was decided that nationalities (and languages) will be eliminated from the list as well, thus allowing for a good number of more useful content words in the list.

Additionally, unlike English, Romanian corpora contain two different spelling systems, the old one and the new one. For instance, the words *întîia* (first) and *întîmpla* (happen) today are spelt as *întâia* and *întâmpla.* It was decided that words using the old spelling would be removed from the final RWL. The reason for this was that learners, once mastered either versions of these words, by pronouncing them, realise instantaneously what the words are. Therefore, either counting or teaching these words separately would be counterproductive. However, some of these are quite high frequency words and thus, the coverage that the list gives can be considered an underestimate. For the current version of the list this issue is believed to have a somewhat minor impact and possibly in future versions they should be addressed accordingly.

Consequently, the final list (the Romanian Word List, RWL), contains the 10,000 most frequent lemmas in Romanian. These single word items have been grouped then into ten

different bands according to their frequency index. Thus, the first 1k band represents the most frequent 1,000 Romanian words and so on.

In order to further analyse the RWL, it is important to discuss the frequency distribution. Numerous studies have shown that in any language highly frequent words will represent a large percentage in any given text (see e.g. Nation 2006). This is because English words such as *the, of, get, give* or Romanian words such as *de, și, eu, vrea* (*from, and, I, want* respectively) are essential function or content words that are needed to formulate meaningful sentences in any subject area. Thus, their occurrence will be high in virtually any corpora.

Table 2 below demonstrates the frequency distribution of the first 5,000 words in the RWL. It can be seen that the first 200 words approximately appear in disproportionally high numbers in comparison to the rest of the words in the list. At around the first 500 words the frequency index stabilises and then gradually decreases. This tendency and the exponential distribution are similar in all languages.

Table 2: Romanian Word Frequency

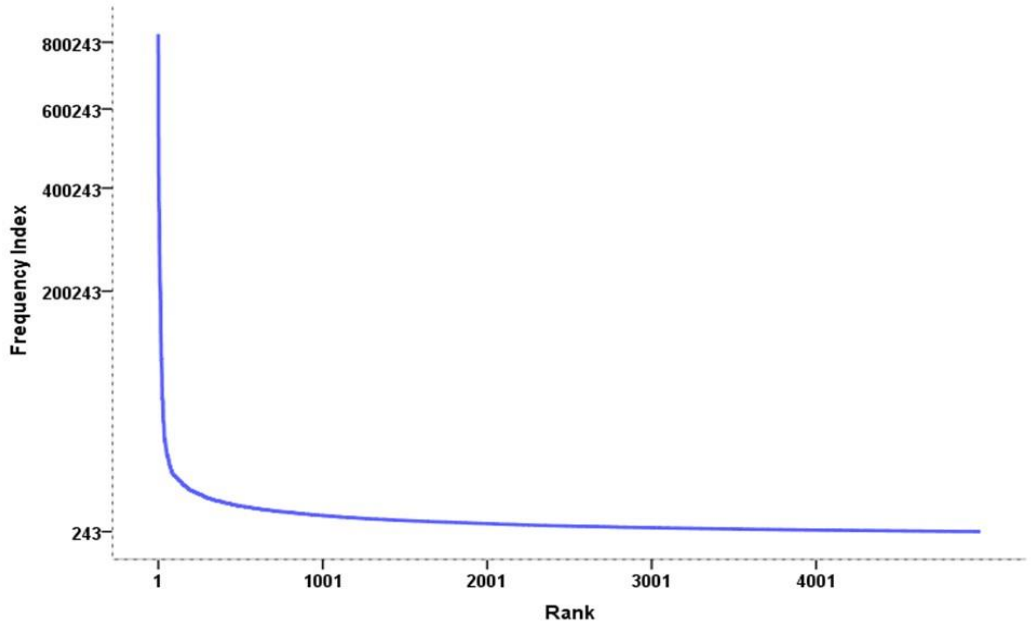

Table 3 aids the comparison between the BNC and the RWL. As we can see the most frequent words (rank 1) in English and Romanian are *the* and *de* and the second ones (rank 2) are *be* and *și* respectively. Their individual frequency number is necessarily dependent on the size of the corpus that has been used. It is interesting to note however that there is a considerable drop between ranks 1 and 2, approximately 2 million words in English and 300k in the case of Romanian. In Romanian this difference becomes more moderate in the case of words with ranks 17 and 18 in contrast to English. Conversely, if words at around the 5k rank are examined, it is visible that the difference between their frequencies is minimised.

Table 3: Frequency comparison between BNC and RWL

Rank	Occurrences		Word		Rank	Occurrences		Word	
	BNC	**RWL**	**BNC**	**RWL**		**BNC**	**RWL**	**BNC**	**RWL**
1	6,187,267	826,777	the	de	**5001**	1188	243	regulatory	card
2	4,239,632	572,717	be	și	**5002**	1188	243	cylinder	jurnalist
17	675,027	138,934	with	pentru	**5017**	1181	242	specialise	mediocru
18	559,596	134,545	do	pe	**5018**	1180	242	steer	nitrat

In line with Nation (2006), this proves that the RWL is properly ordered, as the words' frequency decreases in a similar way to the BNC. To further investigate this, the RWL has to be compared to other word lists and corpora. What is expected is that high frequency words should account for more words in another list or text than words at lower frequency levels.

As has been mentioned, the original list used to construct the RomVLT contained a good number of medical terms. Nevertheless, in order to verify that the two lists share similarities and the word selection criteria used for the RomVLT is reliable, it makes sense to compare the two lists.

Table 4: Original RWL and the RWL

Level	Tokens	Token %	Cumulative %
1	1783	16.65	16.65
2	1048	9.79	26.44
3	985	9.2	35.64
4	962	8.98	44.62
5	900	8.4	53.02
6	817	7.63	60.65
7	699	6.53	67.18
8	592	5.53	72.71
9	427	3.99	76.7
10	321	3	79.7
Not in the lists	2176	20.32	100

As expected, since both lists are extracted from the ROMBAC corpus (except that from the final RWL the medical section has been eliminated) the two lists do indeed share similarities. Overall, the ten frequency bands provide almost 80% of coverage. Furthermore, if the individual frequency levels are taken, it is noticeable that from the 10th level the percentages proliferate up to 16.65 at level 1. This proves that at least in relation to the original version, the RWL is properly distributed.

In order to see how much coverage the bands provide in literature for example, the RWL bands have been run through the literature section of the ROMBAC corpus. It can be noticed (Table 5) that the first 2k most frequent words in Romanian provide over 78% coverage of the almost 7 million running words and if the third band is added, the coverage goes up to 81%. In other words, mastering the first 2,000 to 3,000 words of Romanian would enable learners to demonstrate comprehension of at least four words in every five running words.

This is in line with the English language in which researchers found using various corpora that the most frequent 2,000 words provide around 80% coverage in a variety of contexts, (see Milton 2009; Nation 2006 or Schmitt and Schmitt 2014). Consequently, based on the frequency information extrapolated with the RWL, it is argued here that up to and including the 3,000 most frequent Romanian words should be regarded as high frequency vocabulary and up to and including the 10k level as the Romanian mid-frequency vocabulary. These two lists, as findings suggested provide essential coverage for language learners in most cases. Anything above these levels should be regarded as low frequency vocabulary.

Table 5: RWL Coverage in the Literature section (2)

Level	Token	Token %	Cum Token %
1	4,953,932	71.33	71.33
2	481,144	6.93	78.26
3	255,419	3.68	81.94
4	163,171	2.35	84.29
5	119,148	1.72	86.01
6	96,119	1.38	87.39
7	73,364	1.06	88.45
8	57,972	0.83	89.28
9	46,809	0.67	89.95
10	40,609	0.58	90.53
Not in the list	657,609	9.47	100

Additionally, if lower frequency bands are considered as well (Table 5), it can be observed that the total amount of coverage the 10k most frequent words of Romanian would provide is just about 91%. Turned into comprehension figures, this would still mean that learners on average would encounter one unknown word in every ten running words. On one hand, this could be considered a manageable amount. However learners at this level would still struggle with most ungraded literature texts if only these words were known. On the other hand, as interlanguage develops, learners will incrementally attain a large number of proper nouns, increase their knowledge of compound words and become more capable of recognising and using collocates. Since these have been eliminated from the RWL and it is known that proper nouns are highly frequent especially in the area of literature, the total amount of coverage that the RWL provides might be well over 91%.

There are other issues that have to be considered when compiling frequency lists. The technical advancements available today still make it difficult to distinguish between homonyms, thus words such as *Capitan* (proper noun) and *capitan* are either considered as one word, which in the case of highly frequent words can account for a huge difference, or as separate words, recognized by the capital letter, in which case sentence initial common nouns will be counted as proper nouns.

Nevertheless the RWL is the first Romanian frequency list that is useful and reliable as it not only shows statistical similarities with English, but its practical application seems to be robust as well, as the RomVLT will show. As an initiative, it is believed the RWL will increase the opportunities for further investigations on the subject of Romanian frequency information and lexical knowledge in general.

5. The Romanian Vocabulary Levels Test (RomVLT)

Nation's (1983) Vocabulary Levels Test has been successfully used in an array of studies to investigate vocabulary knowledge quantitatively as it allows for meaningful comparisons between students' overall lexical knowledge, comprehension and foreign language performance. Moreover, it permits effective individual and group comparisons between learners at all levels, even in the case of large number of participants; it can indicate for teachers and researchers where vocabulary teaching should be focused and aids efficient evaluation of progress (see e.g. Molnar 2010; Schmitt, Schmitt and Clapham 2001; Webb and Sasao 2013). Following the compilation criteria of this robust test, the RomVLT also proves to be a valid and reliable vocabulary test and as such, it has been successfully used to compare lexical knowledge in English and Romanian in the case of Hungarian native speakers (Szabo, in preparation).

The VLT and the RomVLT use word frequency information to test receptive vocabulary knowledge. Test items and distractors are selected from five different frequency levels, namely the 2k level (the first 2,000 most frequent words), 3k level, 5k level, the University Word List (UWL) and the 10k level. Since Nation (1983) positions the UWL around the 6k level and as there is no Romanian equivalent to this list (containing high frequency academic or university words), the words for this level were taken from the 6k level of the RWL. Throughout the selection procedure, it has been ensured that noun, verb and adjective clusters match the proportion in the original test at each level. Furthermore, as the aim of the study (Szabo, in preparation) was to test cognate knowledge between English and Romanian, care has been taken that the proportion of cognate words (both test items and distractors) matches in the two tests. Additionally, a small number of words, like *sport* and *trumpet* are cognates with Hungarian. As the purpose was to leave the English test unchanged, the words for the RomVLT were selected in such a way that they contain the same amount of cognates in Hungarian. However, the number of these words is minimal and all are regarded as high frequency words, so these are unlikely have an impact on the overall scores.

Overall, the tests consist of 180 words, featuring five different levels each including six word clusters and each cluster containing six words (36 in total on one level).

Extract from the RomVLT

1 semnătură	
2 supraviețuire	_3_ conversație
3 dialog	_4_ organ al aparatului urinar
4 rinichi	_6_ construcție, edificiu
5 intensitate	
6 clădire	

In order to examine the robustness and usefulness of any test, it is essential to check for its reliability and validity. This is based on Szabo's study (in preparation), which compared Hungarian native speakers' (N = 40) Romanian (L2) and English (L3) lexical knowledge using the VLT and the RomVLT.

The reliability of a test can be investigated in two different ways: the test-retest method or by an equivalence measure, using split-half analysis. As Schmitt (2010: 184) raised several

concerns regarding the test-retest method, the RomVLT's reliability and internal consistency were checked through the split-half analysis. The test was split in two with the 2k, 3k and 5k levels in one half and the 5k, 6k and 10k levels in the other one. The alpha scores for these two parts on individual scores are .924 and .945 respectively ($p < .001$; $N = 40$), where the correlation between the forms equal .821. Cronbach's alpha was calculated for individual scores on the different levels at $\alpha = .937$. Both of these results indicate that the RomVLT seems to be a highly reliable test with a significant level of internal consistency.

The validity of the RomVLT was tested by checking for criterion, content, construct and concurrent validity (see e.g. Schmitt 2010). Criterion validity explores the test design, word selection and the overall procedure that have been followed. Every effort has been made to ensure that the number of levels, words and cognates on the new test matched the VLT. Furthermore, care has been taken that nouns, verbs and adjectives are distributed proportionally on each level. The entire test has been checked by an educated (professor) native speaker of Romanian. It can be thus concluded that the RomVLT is valid in terms of criterion validity.

Construct and content validity are closely associated with each other. Content validity considers the frequency information that the test is using, whereas construct validity questions whether the test is measuring what it purport to measure. In order to investigate these, the results on different levels scored on the RomVLT were considered. It is assumed that high frequency words should be acquired earlier and as lower frequencies are considered, knowledge on these levels should slightly decrease. The following figure illustrates just this.

Figure 1: RomVLT mean results on the different levels

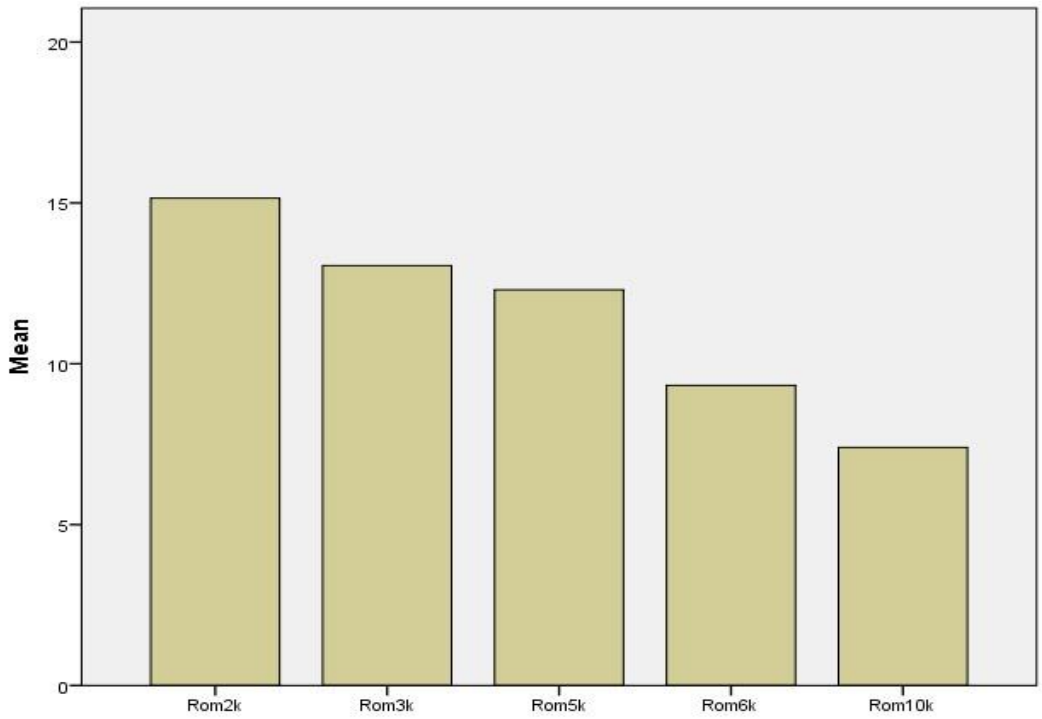

Figure 1 summarises the mean scores on the different frequency levels on the RomVLT. This shows that students reached higher scores on the most frequent Romanian words than on

the less frequent ones. This downward slope indicates that the words chosen from the RWL gradually become more difficult from left to right or from highly frequent to less frequent words. It can be thus assumed that the RomVLT's content validity is of an acceptable level. Furthermore, as the results resemble the theoretical profile of a language learner's lexical knowledge (Meara 1992) it can be asserted that the test's construct validity is satisfied.

Concurrent validity is checked by comparing the test's results to an equivalent test. This is not always straightforward as different tests tap into different type of knowledge by eliciting different type of answers. As the RomVLT has been designed in a similar way to the English VLT, it can be assumed that both measure lexical access in the two different languages. This allowed the two tests to be compared in order to check for concurrent validity. Pearson's correlations between the different levels show statistically significance as follows: for the 2k levels $r = .652$; 3k levels $r = .592$; 5k levels $r = .635$; 6k and UWL level $r = .815$ and the 10k levels $r = .717$ ($p <. 001$, *sig. 2-tailed*, $N = 40$). The overall scores on the two tests show a robust correlation as well: $r = .792$, $p < .001$. The lack of a higher correlation might be due to the fact that lexical knowledge in the two languages cannot be exactly at the same level. Nevertheless, the RomVLT is to be considered valid from this point of view as well.

Based on the above findings, it can be asserted that the RomVLT seems to be a valid and reliable test tool.

6. Conclusion

The aim of this paper was twofold. First, it introduced the RWL, which is the first Romanian Word List that is based on frequency information. It is suggested throughout that as the most frequent 10k words are important for any language learner the RWL should be adopted as 'essential' words for Romanian. The high frequency vocabulary, containing the most frequent 3,000 words of Romanian, provides a large enough coverage for learners to understand texts at beginner level. Furthermore, it was argued that the RWL can be used, adapted and improved for different contexts. It can enable teachers and researchers to establish lexical learning targets for their students. It can aid the design of much needed Romanian graded readers and the methodology can be used for developing specialized word lists for academia or other areas. Using the RWL, reliable and valid testing tools can also be designed, such as the RomVLT.

Second, it introduced a practical application of the RWL, the RomVLT. This proved to be a valid and reliable tool for exploring Romanian learners' vocabulary knowledge on five different frequency bands, allowing for individual and group comparisons, and accounting for cognate knowledge between English and Romanian (Szabo, in preparation). For teachers, this can indicate where vocabulary learning should be focused thus helping them make more informed decisions about the materials they employ and the lexis they consider useful. The test's user-friendliness, and the fact that it assesses language learners' lexical knowledge at different frequency levels, makes it very attractive to teachers and researchers alike.

I believe that the aforementioned developments will broaden the prospects currently viable in the context of teaching Romanian as an additional language and will actively contribute to an understanding of Romanian vocabulary knowledge per se or as compared to English, the spread of vocabulary-based empirical research explorations, the instigation of Romanian specialised lists and last but not least provoke meaningful conversations about current issues in the field.

REFERENCES

Adolphs, S., N. Schmitt, 2003, "Lexical coverage of spoken discourse", *Applied Linguistics* 24, 4, p. 425-38.

Aizawa, K., 2006, "Rethinking frequency markers for English-Japanese dictionaries", in M. Murata, K. Minamide, Y. Tono, Ishikawa S. (eds.) *English Lexicography in Japan*, Tokyo, Taishukan-shoten, p. 108-119.

Arefu, D., 2014, *First Romanian Reader for beginners: bilingual for speakers of English: 1*, Language Practice Publishing.

Axenti, V., M. Verșina, 2009, *Metodica predării limbii și literaturii române (în gimnaziu și liceu): Suport de curs*, Cahul: USC.

Bârlea, P. G., M. Cerkez, 2005, *Limba româna. Fonetică și vocabular, manual pentru programul PIR*, Ministerul Educației și Cercetării.

Biriș, G., D.V. Burlacu, E. Șoșa, 2011, *Antonime. Sinonime. Analogii. Vocabular minimal al limbii române (cu traducere în limba engleză)*, București, Saeculum I.O.

Brezina, V. & Gablasova, D., 2015, "Is there a core general vocabulary? Introducing the New General Service List", *Applied Linguistics*. 36, 1, p. 1-22.

Coxhead, A., 2000, "A new academic word list", *TESOL Quarterly*, 34, p. 213-239.

Davies, M,. 2008, "The Corpus of Contemporary American English – a Useful Tool for English Teaching and Research", *Computer-Assisted Foreign Language Education in China*, 5, p. 24-31.

Dina, A. T., 2013, "Successful Approach for Teaching Romanian as a Foreign Language", *Procedia – Social and Behavioral Sciences*, 70, p. 1032-1037.

Erten, I. H., M. Tekin, 2008, "Effects of vocabulary acquisition of presenting new words in semantic sets versus semantically unrelated sets", *System*, 36, 3, p. 407-422.

Gardener, D., M. Davies, 2013, "A New Academic Vocabulary List", *Applied Linguistics*, 35, p. 1-24.

Golu, P., M. Zlate, E. Verza, 1992, *Psihologia Copilului, Manual pentru a clasa a XI-a, școli normale*, Editura Didactică și Pedagogică, București.

Ion, R., E. Irimia, D. Ștefănescu, D. Tufiș., 2012, "ROMBAC: The Romanian Balanced Annotated Corpus", in *8th International Conference on Language Resources and Evaluation*, p. 339-344.

Kilgarriff, A., Rychly, P., Smrž P., Tugwell, D., 2004, "The Sketch Engine", In Proceedings of *Euralex 2004*, Lorient, France, p. 105–116.

Lessard-Clouston, M., 2012, "Vocabulary Learning and Teaching: Pedagogy, Research, and Resources", in *9th Christians in English Language Teaching (CELT 2012) Conference, Teaching With Excellence Strand*, Chinese University of Hong Kong, Hong Kong, China.

Macoveiciuc, M., A. Kilgarriff, A., 2010, "The RoWaC Corpus and Romanian Word Sketches", in D. Tufiș, C. Forăscu (eds.), *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*, Romanian Academy Publishing House, Bucharest.

Meara, P. M., 1980, "Vocabulary Acquisition: A Neglected Aspect of Language Learning", *Language Teaching*, 13, p. 221–246.

Meara, P. M., 1992, EFL Vocabulary Tests. University College Swansea: Centre for Applied Language Studies.

Milton, J., 2009, *Measuring second language vocabulary acquisition*, Bristol, UK, Multilingual Matters.

Miralpeix, I., 2008, *The influence of age on vocabulary acquisition in English as a Foreign Language*. PhD thesis, Universitat de Barcelona.

Molnar, T., 2010, "Cognate Recognition and L3 Vocabulary Acquisition", *Acta Universitatis Sapientiae, Philologica*, 2, 337−349.
Nation, P. and Coxhead, A., 2014, "Vocabulary size research at Victoria University of Wellington, New Zealand", *Language Teaching*, 47, 3, p. 398 – 403.
Nation, P., P. Meara, 2010, "Vocabulary", in N. Schmitt (ed.) *An introduction to applied linguistics* (2nd ed.), London, Hodder Education, p. 252-267.
Nation, P., S. Webb, 2011, *Researching and analyzing vocabulary*. Boston, MA, Heinle.
Nation, P., 1983, "Testing and teaching vocabulary", *Guidelines*, 5, p. 12-25.
Nation, P., 2006, "How Large a Vocabulary is Needed for Reading and Listening?", *The Canadian Modern Language Review*, 63, 1, p. 59–81.
Norel, M., L. Pop, 2005, *Limba română ca a doua limbă*. Ministerul Educației și Cercetării. Unitatea de Management a Proiectului pentru Învțământul Rural, București.
O'Loughlin, R., 2012, "Tuning In to Vocabulary Frequency in Coursebooks", *RELC Journal*, 43, 2, p. 255-269.
Platon, E., D. Burlacu, I. Sonea, I., 2011, *Procesul de predare-învățare a limbii române ca limbă nematernă (RLNM) la ciclul primar/gimnazial/liceal*, RLNM: P1-P6 (6 volume), ClujNapoca, Casa Cărții de Știință.
Schmitt, N., Schmitt, D. and Clapham, C., 2001, "Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test", *Language Testing* 18, 55-88.
Schmitt, N., D. Schmitt, 2014, "A reassessment of frequency and vocabulary size in L2 vocabulary teaching", *Language Teaching*, 47, 4, 2014, p. 484-503.
Schmitt, N., 2010, *Researching Vocabulary: A Vocabulary Research Manual*, Palgrave Press.
Sîrghie, A., 2009, *Metodica predării limbii și literaturii române în învățământul preșcolar și primar*, Sibiu, Alma Mater.
Szabo, Cs., in preparation, "*Exploring L2 and L3 vocabulary size and the effect of cognate instruction".*
Tinkham, T., 1993, "The effect of semantic clustering on the learning of second language vocabulary", *System*, 21, 3, p. 371-380.
Treffers-Daller, J., J. Milton, 2013, "Vocabulary size revisited: the link between vocabulary size and academic achievement", *Applied Linguistics Review*, 4, 1, p. 151–172.
Webb, S. A., Y. Sasao, 2013, "New Directions in Vocabulary Testing", *RELC Journal*, 44, p. 263-277.
West, M., 1953, *A General Service List of English Words*, London, Longman, Green & Co.
Wilkins, D. A., 1974, *Linguistics in Language Teaching*, London, Edward Arnold.